Jesan Bapari Akter
201496637

COMP 341, report on M-CNN paper:

	Semantic Segmentation	Object Detection	Instance Segmentation	2D Grasp Detection	6DOF Pose Estimation
Task Description	Each pixel in the image is assigned a semantic label	Objects within an image are identified and localised	Objects within an image are identified and segmented	Identifies suitable locations on an objects surface where a hand can grasp it	Determining the position and orientation of object in a 3-dimensional space
Output	Segmented image where each pixel is assigned a class label corresponding to the region it belongs to	Bounding boxes (defined by its coordinates) encompassing the detected objects within an image, as well as class labels indicating type of object detected	Segmented masks outline individual objects within an image, as well labels indicating the category of each object	Annotations indicating viable grasping positions on objects within an image. May include probabilities associated with each grasp	Estimated parameters defining the position and orientation of object, includes the translation vector and the rotation matrix. May also include confidence scores linked with estimated pose
Loss	Cross-entropy	Anchor-based (Smooth L1) or region-based	Combined loss (e.g. bounding box regression, mask binary cross-entropy	Intersection over Union, Grasp loss	Regression loss (e.g.L2, smooth L1)
Key datasets	Cityscapes, ADE20K, PASCAL VOC	MS- COCO, Open Images, Pascal VOC	MS-COCO, Cityscapes	YCB-Video Dataset, Cornell Grasping Dataset	LineMOD, T-LESS, YCB-Video Dataset
Key papers	Deeplab, FCN, U-Net	YOLO, SSD, Faster R-CNN	Mask-RCNN, PANet, FCIS	Retina Grasp, Dex-Net, GraspNet	PoseCNN, DeepIM, PVNet

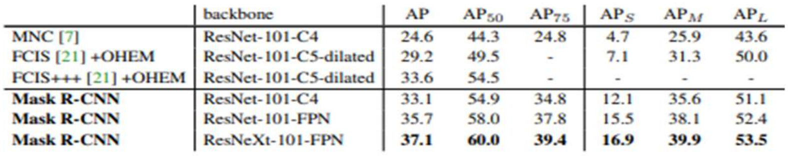
Jesan Bapari Akter
201496637

Mask-CNN made a breakthrough in the field of computer vision, especially in the field of computer vision. This paper is an extension to the Faster R-CNN, where it's able to perform object detection and instance segmentation simultaneously. The addition of a mask prediction branch into Faster R-CNN and its ROIAlign Layer to resolve misalignment issues, Mask R-CNN was able to achieve amazing results in various datasets.

This paper discusses the evolution of object detection and instance segmentation methods. Initially, R-CNN valuated convolutional networks independently on each region of interest, with extension to improve speed and accuracy, Faster R-CNN improved this further with a Region Proposal Network. For instance, segmentation, earlier approaches relied on bottom-up segments, but these methods were slow and accuracy was low. However, the method proposed by the paper emphasises parallel prediction of masks and class labels, which made it simple and flexible.

The addition of the third branch for object mask prediction alongside class labels and bounding-box offsets enabled better spatial layout extraction. Mask R-CNN differs in predicting masks independently of class labels, to enhance segmentation accuracy, decoupling masks and class prediction is done, this involves a multi-task loss function. RoIAlign layer is introduced to preserve pixel-to-pixel alignment, employing bilinear interpolation for accurate feature extraction and avoiding quantization of RoI boundaries. Various architectures, such as ResNet and FPN tests the model's flexibility and implementing interface and training strategies consistent with Fast R_CNN work. Instance segmentation is challenging because it requires correct detection of all objects in the image while also precisely segmenting each instance., therefore this is a promising solution that improves speed and accuracy.

Experiments were conducted using Mask R-CNN, it outperformed previous methods such as FCIS and MNC, on the COCO dataset, without additional techniques like multi-scale training



	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
MNC [7]	ResNet-101-C4	24.6	44.3	24.8	4.7	25.9	43.6
FCIS [21] +OHEM	ResNet-101-C5-dilated	29.2	49.5	-	7.1	31.3	50.0
FCIS+++ [21] +OHEM	ResNet-101-C5-dilated	33.6	54.5	-	-	-	-
Mask R-CNN	ResNet-101-C4	33.1	54.9	34.8	12.1	35.6	51.1
Mask R-CNN	ResNet-101-FPN	35.7	58.0	37.8	15.5	38.1	52.4
Mask R-CNN	ResNeXt-101-FPN	**37.1**	**60.0**	**39.4**	**16.9**	**39.9**	**53.5**

AP: averaged over IoU thresholds

Loads of aspects were analysed of Mask R-CNN , including RoIAlign layer, mask branch, multinomial vs independent masks and backbone architecture, showing improvements in performance. Also, performed well in object detection, performing better than previous state-of-the-art models, there was significant improvement when using ResNeXt-101-FPN backbone, with a margin of 3.0 points box AP over the best previous. Training time is fast and interface time is around 195ms per image, which is suitable for practical use.

Paper goes into human pose estimation, predicting masks for each key point and treating key points as one-hot masks. Experiments outperform previous methods in speed and simplicity while achieving amazing accuracy. Multi-task learning indicates improvements with segments, key points and unified model predicting boxes at the same time. Furthermore, the effectiveness of RoIAlign over RolPool is shown for key point detection, which shows the significance of accurate localization.

The paper discusses Mask R-CNN really well, pointing out its strengths such as performance, flexibility, innovation etc.., discussing previous model. However, it fails to address some challenges and future work that it may require such as hard to scale with large datasets, better optimization is required without sacrificing accuracy, enhancing robustness.